UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: THE NEW YORK TIMES COMPANY v. MICROSOFT CORPORATION, et al., No. 23-cv11195	25-md-3143 (SHS) (OTW) Hon. Sidney H. Stein Hon. Ona T. Wang

[PROPOSED] ORDER RE: CONVERSATION DATA SAMPLING

WHEREAS, the Court has presented the "hypothetical[]" that "ChatGPT user[s]" who use ChatGPT to "get around [a] pay wall" might delete their conversations to avoid potential liability, 1/22/2025 Hr'g Tr. at 38:14–20, and that this behavior may result in a material difference between the conversation data OpenAI has retained and the conversation data that OpenAI does not retain on account of its commitment to honor user-initiated deletions, MDL ECF 42 at 5.

WHEREAS, the Court on May 16, 2025 requested that the parties propose a "means to determine" whether there is such a difference that might be relevant to this litigation. *Id.* at 5.

WHEREAS, the Court subsequently endorsed OpenAI's[1] proposal to undertake "some form of sampling that enables us to compare whether there's a . . . difference in the hit rate in these two different pools of data," 5/27/2025 Afternoon Hr'g Tr. at 43:11–18, and confirmed that the Court was "absolutely not . . . interested in" ordering OpenAI to "produce[] to the plaintiffs" any data that OpenAI began retaining in compliance with the Court's May 13, 2025 order, *id.* at 3:18–25; and

[1] "OpenAI" shall refer to Defendants OpenAI, Inc., OpenAI OpCo, LLC, OpenAI GP, LLC, OpenAI, LLC, OpenAI OpCo, LLC, OpenAI Global LLC, OAI Corporation, and OpenAI Holdings, LLC. "News Plaintiffs" shall collectively refer to Plaintiffs The New York Times Company; Daily News, LP; Chicago Tribune Company, LLC; Orlando Sentinel Communications Company, LLC; Sun-Sentinel Company, LLC; San Jose Mercury-News, LLC; DP Media Network, LLC; ORB Publishing, LLC; Northwest Publications, LLC; and the Center for Investigative Reporting.

WHEREAS, on May 29, 2025, the Court ordered the parties to begin sampling the "30-day tables of consumer output log data" and ordered OpenAI to submit a proposed order regarding the sampling process, consistent with the discussion at the May 27 hearing, MDL ECF 79 ¶ 2;

ORDERED, that the following procedures shall apply to the sampling of the consumer output log data set forth in the Court's May 29, 2025 Order:

1. As soon as practicable following entry of this Order, OpenAI shall generate two data samples using a simple random sampling methodology as follows:

    a. Each sample shall consist of the contents of individual rows in the table of consumer output log data that OpenAI has segregated and preserved in response to the Court's May 13, 2025 Order, MDL ECF No. 33 (the "Preservation Order"), where each row contains, *inter alia*, individual text prompts for requests generated by users whom OpenAI reasonably believes are located in the United States and corresponding text outputs generated by ChatGPT Free, Plus, or Pro for the models at issue (*see Authors Guild* ECF No. 293) between April 14, 2025 and May 14, 2025 (hereinafter, "Conversation Data").

    b. OpenAI shall use best efforts to draw independent and identically distributed samples from the sample populations described below:

        i. The sample population for the first sample ("Test Sample") shall be all Conversation Data.

        ii. The sample population for the second sample ("Control Sample") shall be all Conversation Data that (i) was not generated through the "Temporary Chat" feature and (ii) is not subject to a user-initiated deletion. *See* MDL ECF 66 ¶ 3.

    c. The size of the data samples shall be determined according to the formula and parameters adopted by the court in *Concord Music Grp., Inc. v. Anthropic PBC*, Dkt. 377, Case No. 5:24-cv-03811-EKL (SVK) (N.D. Cal. May 23, 2025), thus resulting in a sample size for each of the Control Sample and the Test Sample of five (5) million rows of Conversation Data, which amounts to ten (10) million rows of Conversation Data across both samples.

    d. OpenAI shall undertake best efforts to anonymize the data in the Control Sample and Test Sample prior to any use of the data pursuant to this Protocol.

2. Within seven (7) days of generating the Test Sample and Control Sample, OpenAI shall run keyword searches across prompts in Conversation Data in the Test Sample and Control Sample, using search terms (attached hereto as Appendix A) comprised of:

a. the names of News Plaintiffs' publications;

b. the News Plaintiffs' domains; and

c. "paywall" or "pay wall."

3. After running the searches in the foregoing paragraph, OpenAI will provide News Plaintiffs with two tables (in a native spreadsheet format) containing hit counts for each search term. One table shall contain corresponding hit counts for the Control Sample and the other table shall contain corresponding hit counts for the Test Sample.

4. All data and analyses derived from the samples, including hit count tables, shall be designated as HIGHLY CONFIDENTIAL – OUTSIDE COUNSEL'S EYES ONLY pursuant to the operative Protective Order in this action.

5. All data and analyses derived from the samples, including hit count tables, may only be used for the sole purpose of determining the existence of any material difference between the Control Sample and the Test Sample, as contemplated in the May 29, 2025 Order.

6. Seven (7) days after OpenAI provides the hit count tables for the Test and Control Samples described above, OpenAI and News Plaintiffs shall submit a joint letter of no more than three (3) pages (no more than one-and-a-half (1.5) pages per side) that attaches the hit count tables and presents each side's position regarding whether there exists any material difference in hit counts between the two samples.

7. If News Plaintiffs' findings submitted in paragraph 6 above do not purport to show the existence of a material difference between the Test and Control Samples, the Preservation Order shall be vacated. If, however, OpenAI and News Plaintiffs submit disputed findings as to the existence of a material difference between the Test and Control Samples, and that dispute is resolved by the Court in OpenAI's favor, the Preservation Order shall thereafter be vacated, and OpenAI may recover any costs associated with its retention of the consumer output log data that OpenAI has segregated and preserved in response to the Preservation Order that it incurred from the date of the joint letter in paragraph 6 above.

Dated: June ___, 2025 **SO ORDERED.**

HON. ONA T. WANG
United States Magistrate Judge